Estimation of the local diffusion tensor and normalization for heterogeneous correlation modelling using a diffusion equation

Olivier Pannekoucke, S. Massart

QUARTERLY JOURNAL OF THE ROYAL METEOROLOGICAL SOCIETY
Q. J. R. Meteorol. Soc. **00**: 1–14 (2008)
Published online in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/qj.000

Estimation of the local diffusion tensor and normalization for heterogeneous correlation modelling using a diffusion equation

O. Pannekoucke*[1] and S. Massart[2]
[1] *Météo-France, CNRM-GAME, Toulouse and* [2] *CERFACS*

Abstract: As the background error covariance matrix is a key component of any assimilation system, its modelling is an important step. Usually, this matrix is decomposed into correlations and variance matrices. An interesting method for modelling the correlation matrix of the background error for complex geometry, like ocean grid, consists in computing correlation functions using a diffusion operator. The background error correlation functions can be estimated for example from an ensemble of perturbed forecasts. The diffusion operator is able to represent heterogeneous correlation functions at a reasonable numerical cost. But a first challenge resides in the determination of the local diffusion tensor corresponding to the local correlation function. Then the second challenge resides in the determination of the normalization to make sure that the matrix modelled through the diffusion operator is a correlation matrix. In this article, we propose to build a background error correlation matrix using a diffusion operator based on a local diffusion tensor. The estimation of this local tensor is performed using an ensemble of perturbed forecasts. A validation within a randomization method illustrates the feasibility and the accuracy of the proposed method. In particular, it is shown that the local geographical variations of diagnosed correlation functions (through an ensemble of perturbed forecast) are well represented.

This is first illustrated in an analytical one-dimensional framework. In that context, the diffusion field and the normalization field are deduced from a given correlation length-scale field. The resulting length-scales are shown to correspond to the initial length-scale when the given length-scale field spectrum is red. The approximate normalization, computed from the local length-scale, is close to the true normalization under the same condition of a red spectrum.

Then, the method is illustrated in a real context using an ensemble of perturbed forecasts from the MOCAGE-PALM assimilation system. In that case, length-scale and anisotropy diagnosis reveal the complexity of the correlation of stratospheric ozone forecast errors. The local diffusion tensor deduced from these diagnosis are shown be able to represent such an existing heterogeneity and anisotropy. As in the one-dimensional case, the approximate normalization, based on the local diffusion tensor, appears to be a really good approximation of the true normalization.

WARNING : This is a preprint of an article accepted for publication in QUARTERLY JOURNAL OF THE ROYAL METEOROLOGICAL SOCIETY *Q. J. R. Meteorol. Soc.* **134**:1425–1438 (2008) see the website for final version http://www.interscience.wiley.com/

KEY WORDS background-error correlation modelling; diffusion operator; assimilation; ensemble

Received 1 February 2008; Revised 11 June 2008; Accepted 13 June 2008

1 Introduction

The main aim of atmospheric or oceanographic data assimilation is to construct the most likely representation of the flow from available observations. However, the lack of observations compared to the number of degree of freedom of numerical flow involves the use of a prediction correction approach. The problem is thus solved when, as a prior information, the background is introduced. In practice, the background equals the latest available forecast. The prediction correction approach commonly introduces some new unknown quantity: the observation error covariance matrix and the background error covariance matrix.

The paper focuses on the background error covariance matrix. A first difficulty comes from the estimation of the correlation part of the covariance matrix: it is hard to estimate such a matrix because of its huge size and because the estimation is affected by sampling noise. Then, the matrix generally has to be modelled because it is extremely costly to compute it and to store it under a matrix form.

Some models of the background correlation matrix are based on a diagonal assumption in spectral space (Courtier *et al.*, 1998) leading to homogeneous covariances. More recently, the diagonal assumption in the wavelet space has been introduced in order to model

*Correspondence to: Météo-France CNRM/GMAP/RECYF, 42 av. G. Coriolis, 31057 Toulouse Cedex France. e-mail: olivier.pannekoucke@meteo.fr

some heterogeneous correlations (Fisher, 2003). This last method offers some interesting filtering properties: using these techniques back to spatially averaging the local correlation functions (Pannekoucke *et al.*, 2007).

Another possible model of the background covariance matrix has been proposed by Weaver and Courtier (2001). It is based on the use of a generalized diffusion equation which aims to construct heterogeneous correlations. This approach allows to easily constructs Gaussian-like correlation functions even over a complex domain *e.g.* in ocean modelling where the cost involve tortuous boundaries constraints. The heterogeneity comes from the variation over the domain of the local diffusion tensor. However, there is no obvious way to obtain such a local diffusion tensor. Moreover, the model based on the diffusion operator does not lead to a valid correlation matrix and has to be normalized in order to ensure a local variance equal to unity.

Beyond the representation of the heterogeneity, the diffusion equation model is able to represent anisotropy components of the background error correlation function. This is achieved when the diffusion tensor is not diagonal. Recently, Liu *et al.*(2007) have shown, by a tricky way, that it is possible to represent flow-dependent anisotropy in recursive filters. In their study, they have illustrated that the anisotropic filters improves the analysis of fine-scale structures. It has to be noted that some alternative ways to construct model diffusion based on Gaussian correlation exists, for instance, the use of an iterated Laplacian method (Derber and Rosati, 1989; Egbert *et al.*, 1994) or the recursive filter approach (Purser *et al.*, 2003 a&b).

This article adresses the issue of the estimation of the local diffusion tensor and of the local normalization. Only a simple diffusion operator is considered (and not the generalized diffusion equation). This estimation is partly based on the computation of the local length-scale (Belo Pereira and Berre, 2006; Pannekoucke *et al.*, 2008).

For real applications, the length-scale can be estimated from an ensemble of perturbed forecasts (Houtekamer *et al.*, 1996; Fisher, 2003). The main feature of the ensemble method is to provide flow-dependent information about the spreading of background error in the background phase space at a given date. Even if such an ensemble features the error of the day (Kalnay, 2002), perturbed ensembles over a period of several days are often used. In this way, a large ensemble is available and gives access to the mean spreading of background error, with a reduction of the sampling noise. A large ensemble serves to calibrate static components of background covariance model (Fisher, 2003). This paper focuses on the representation of static geographical variations of background correlation functions. This is a preliminary step towards flow-dependent background correlation modelling.

The structure of the paper is as follow. In section 2, basics of data assimilation are recalled with an emphasis on correlation modelling using a diffusion equation. Section 3 explains how to estimate the local diffusion tensor and the local normalization. This method is applied in section 4, for the particular case of an analytical one dimensional circular framework. Finally, in section 5, a real application is proposed in order to illustrate the method for the particular case of the MOCAGE-PALM chemical global assimilation system. Conclusions are given in section 6.

2 About data assimilation

2.1 Variational data assimilation

Data assimilation consists in finding the more likely state $\boldsymbol{x}^a$ (the analysis) of the atmosphere or ocean, knowing a background state $\boldsymbol{x}^b$ and observations $\boldsymbol{y}^o$. Compared to the true state $\boldsymbol{x}^t$ (not known in practice), the background corresponds to the truth plus a background error $\boldsymbol{\varepsilon}^b$ so that $\boldsymbol{x}^b = \boldsymbol{x}^t + \boldsymbol{\varepsilon}^b$. Similarly, an error $\boldsymbol{\varepsilon}^o$ occurs on the observation state so that $\boldsymbol{y}^o = \mathbf{H}\boldsymbol{x}^t + \boldsymbol{\varepsilon}^o$, where $\mathbf{H}$ is the (linear in this reminders) observation operator that maps the model space into the observations space. Both errors are assumed to be uncorrelated *i.e.* $\mathbb{E}\left(\boldsymbol{\varepsilon}^o {\boldsymbol{\varepsilon}^b}^T\right) = \mathbf{0}$, where $\mathbb{E}$ denotes the expectation function and ^{T} denotes the transposition operation. It is also assumed that $\mathbb{E}\left(\boldsymbol{\varepsilon}^o\right) = 0$ and $\mathbb{E}\left(\boldsymbol{\varepsilon}^b\right) = 0$. The analysis is sought as beeing a corrected state $\boldsymbol{x}^a = \boldsymbol{x}^b + \delta\boldsymbol{x}^a$ of the background state $\boldsymbol{x}^b$, where $\delta\boldsymbol{x}^a$ is the increment to be estimated. The variational approach for resolving the issue consists in minimizing the cost function

$$\mathcal{J}(\boldsymbol{v}) = \boldsymbol{v}^T\boldsymbol{v} + \left(\boldsymbol{d} - \mathbf{H}\mathbf{B}^{1/2}\boldsymbol{v}\right)^T \mathbf{R}^{-1} \left(\boldsymbol{d} - \mathbf{H}\mathbf{B}^{1/2}\boldsymbol{v}\right),$$

where $\mathbf{B} = \mathbb{E}\left(\boldsymbol{\varepsilon}^b {\boldsymbol{\varepsilon}^b}^T\right)$ is the background error covariance matrix, $\mathbf{R} = \mathbb{E}\left(\boldsymbol{\varepsilon}^o {\boldsymbol{\varepsilon}^o}^T\right)$ is the observation error covariance matrix, and $\boldsymbol{d} = \boldsymbol{y}^o - \mathbf{H}\boldsymbol{x}^b$ is the misfit or the innovation vector. The suitable change of variable $\boldsymbol{v} = \mathbf{B}^{-1/2}\delta\boldsymbol{x}$ is used in order to improve the conditioning of the minimizing problem. The square-root matrix $\mathbf{B}^{1/2}$ of the covariance matrix $\mathbf{B}$ is defined so that

$$\mathbf{B} = \mathbf{B}^{1/2}\mathbf{B}^{T/2}.$$

The matrix $\mathbf{B}$ plays a key role in data assimilation scheme as it contributes to filter the observational error and it spreads spacially the correction provided by the innovation. Due to its huge size and to the difficulties to estimate it, $\mathbf{B}$ is often modelled. When it is physically acceptable, $\mathbf{B}$ is expanded as the product

$$\mathbf{B} = \mathbf{\Sigma}\mathbf{C}\mathbf{\Sigma}^T,$$

where $\mathbf{\Sigma}$ corresponds to the diagonal matrix of standard deviations and $\mathbf{C}$ is the correlation matrix. As the variational cost function $\mathcal{J}$ requires $\mathbf{B}^{1/2}$, this matrix is formulated according to the former decomposition of $\mathbf{B}$, as

$$\mathbf{B}^{1/2} = \mathbf{\Sigma}\mathbf{C}^{1/2}, \tag{1}$$

where $\mathbf{C}^{1/2}$ is a square-root matrix of $\mathbf{C}$.

One classic model of $\mathbf{B}$ is to assume that the correlation functions are Gaussian. This is a first approximation of correlation functions that is not alway appropriate (Lorenc, 1992), but that can be considered as building bricks to model much more sophisticated correlation structures (Gneiting, 1999b; Purser *et al.*, 2003 a & b). Moreover, the associated symmetric matrix $\mathbf{C}$ is positive-definite. This shape of correlation function is realistic and it compensates the lack of knowledge about the true statistics. Note that for spherical domains, the chordal distance has to be used in order to obtain a valid (positive definite) correlation function (Gaspari and Cohn, 1999; Weber and Talkner, 1993; Gneiting, 1999 a&b). On the sphere (also the torus and the circle) such a Gaussian correlation is easily constructed as it leads to a diagonal operator in spectral space. But this is no longer the case for complex geometry as encountered in ocean modelling (or atmospheric Limited Area Model) with irregular coastline boundaries. Furthermore, the way the $\mathbf{B}$ operator is modelled must be efficient in term of computational cost in order to avoid to penalize real-time applications. Thus, low cost strategies have to be found to model the covariance matrix $\mathbf{B}$, providing its square root $\mathbf{B}^{1/2}$ for variational purposes.

2.2 Covariance modelling with diffusion operator

A diffusion equation applied on the η variable has the general form

$$\partial_t \eta = \boldsymbol{\nabla} \cdot (\boldsymbol{\nu} \boldsymbol{\nabla} \eta)\,, \tag{2}$$

where $\boldsymbol{\nu}$ is the local diffusion tensor. The equation is assumed to occur over a particular manifold $\mathcal{D}$ associated with a measure $d\omega$. Let $\boldsymbol{x}$ denotes a particular point of $\mathcal{D}$. For the sake of simplicity, one can imagine $\mathcal{D}$ as $\mathbb{R}^n$. $\mathcal{D}$ can also be a numerical model area with its particular mesh such as an homogeneous sphere (representation in spherical harmonic with triangular truncation), or like an ocean grid.

The general solution of the linear Eq.(2) is given as

$$\eta(\boldsymbol{x}, t) = \int_{\mathcal{D}} G^{\boldsymbol{\nu},t}(\boldsymbol{x}, \boldsymbol{x}') \eta(\boldsymbol{x}', t=0) d\omega(\boldsymbol{x}'), \tag{3}$$

where the kernel $G^{\boldsymbol{\nu},t}$ is the Green function, solution at time t of Eq.(2) for the particular initial condition

$$\eta(\boldsymbol{x}', t=0) = \delta_{\boldsymbol{x}}(\boldsymbol{x}'),$$

with $\delta_{\boldsymbol{x}}$ the distribution of Dirac located at point $\boldsymbol{x}$. The Dirac distribution plays a key role in the resolution of the diffusion equation. This distribution is defined so that for all numerical function f over $\mathcal{D}$,

$$\int_{\mathcal{D}} f(\boldsymbol{x}') \delta_{\boldsymbol{x}}(\boldsymbol{x}') d\omega(\boldsymbol{x}') = f(\boldsymbol{x}).$$

Note that the definition of $\delta_{\boldsymbol{x}}$ clearly depends on the measure $d\omega$ over $\mathcal{D}$. It also depends on the vector space in which the solution is sought.

The idea developed by Weaver and Courtier (2001), was to take advantage of such a solution. They have proposed to model the background correlation as the kernel $G^{\boldsymbol{\nu},t}$ and they have proven it to correspond to a correlation tensor (after an appropriate normalization). The correlation function at a given position $\boldsymbol{x}$ is thus defined as $G^{\boldsymbol{\nu},t}(\cdot, \boldsymbol{x})$. This correlation function is constructed as the integration of the non-constant diffusion equation with the initial condition $\delta_{\boldsymbol{x}}$.

Following Weaver and Courtier (2001), the practical integration is achieved as a product of $\mathbf{L}$ by $\mathbf{W}^{-1}$ where the linear operator $\mathbf{L}$ corresponds to the time integration of Eq.(2) over the period $[0, t]$ and where $\mathbf{W}^{-1}$ is the inverse metric product. This latest operator $\mathbf{W}^{-1}$ is defined as follow. Let $\chi_{\boldsymbol{x}}$ denotes the characteristic function related to a point $\boldsymbol{x}$ defined by

$$\chi_{\boldsymbol{x}}(\boldsymbol{x}') = \begin{cases} 0 & \text{if } \boldsymbol{x}' \neq \boldsymbol{x}, \\ 1 & \text{if } \boldsymbol{x}' = \boldsymbol{x}. \end{cases}$$

The inverse metric tensor is formally defined as the linear operator $\mathbf{W}^{-1}$ such as

$$\mathbf{W}^{-1}(\chi_{\boldsymbol{x}}) = \delta_{\boldsymbol{x}}. \tag{4}$$

Note that $\mathbf{W}^{-1}$ depends on the measure $d\omega$ as $\delta_{\boldsymbol{x}}$ depends on the measure and on the vector space where the solution of the diffusion equation is sought. Furthermore, the definition of $\mathbf{W}^{-1}$ can be restricted to a particular functional subspace where the solution is sought. In that particular case, χ has to be projected in functional subspace: on the homogeneous circle at truncation T, one has to consider the projection of χ on the discretized circle generated by $2T+1$ exponential functions (see appendix B).

Then the correlation tensor is defined as $\tilde{\mathbf{C}} = \mathbf{L}\mathbf{W}^{-1}$. Weaver and Courtier (2001) have shown that this tensor can be also formulated as

$$\tilde{\mathbf{C}} = \mathbf{L}^{1/2} \mathbf{W}^{-1} \mathbf{L}^{T/2}, \tag{5}$$

where $\mathbf{L}^{1/2}$ is the half time integration or the propagator from initial time to time $t/2$. However, the standard deviations modelled by $\tilde{\mathbf{C}}$ are not equal to unity and to obtain a correlation tensor, $\tilde{\mathbf{C}}$ is normalized as

$$\mathbf{C} = \boldsymbol{\Lambda} \tilde{\mathbf{C}} \boldsymbol{\Lambda}^T$$

where $\boldsymbol{\Lambda}$ is the diagonal tensor of inverse standard deviation of $\tilde{\mathbf{C}}$. The correlation tensor can be expanded as

$$\begin{aligned} \mathbf{C} &= \left(\boldsymbol{\Lambda} \mathbf{L}^{1/2} \mathbf{W}^{-1/2}\right) \left(\boldsymbol{\Lambda} \mathbf{L}^{1/2} \mathbf{W}^{-1/2}\right)^T, \\ &= \mathbf{C}^{1/2} \mathbf{C}^{T/2}. \end{aligned} \tag{6}$$

Equations Eq.(1), (5) and (6) lead to a practical formulation in variational data assimilation scheme of the square root background covariance error matrix with $\mathbf{B}^{1/2} = \boldsymbol{\Sigma} \boldsymbol{\Lambda} \boldsymbol{L}^{1/2} \mathbf{W}^{-1/2}$. This covariance modelling based on the diffusion equation is particularly appropriate in complex geometry or to model heterogeneous background matrix. The cost of the formulation is mainly due to the cost of the propagator $\mathbf{L}$.

The next subsection describes this expansion in the particular case of the real plane.

2.3 Correlation modelling with constant diffusion equation on the real plane

On the real plane $\mathbb{R}^2$, a point is denoted by $\boldsymbol{x} = (x, y)$, the two dimensional diffusion equation is defined by Eq.(2) with $\boldsymbol{\nu}$ the field of the diffusion tensor. $\boldsymbol{\nu}(\boldsymbol{x})$ corresponds to the local diffusion tensor that *a priori* depends on the position of $\boldsymbol{x}$. This local tensor can be expanded as

$$\boldsymbol{\nu}(\boldsymbol{x}) = \begin{pmatrix} \nu_x(\boldsymbol{x}) & \nu_{x,y}(\boldsymbol{x}) \\ \nu_{x,y}(\boldsymbol{x}) & \nu_y(\boldsymbol{x}) \end{pmatrix}. \tag{7}$$

Note that the diffusion tensor is assumed symmetric (this assumption is not necessary in constant diffusion tensor case, as $\boldsymbol{\nu}$ and $\boldsymbol{\nu}^T$ lead to the same Laplacian operator even when $\nu_{x,y} \neq \nu_{y,x}$).

As seen in the previous section, the operator $\mathbf{L}$ corresponds to the solution Eq.(3). If $G^{\boldsymbol{\nu},t}$ is analytically known, then it determines the expansion Eq.(5). In the particular $\boldsymbol{\nu}-$constant case, an analytical solution can be derived. It can be shown (see appendix A) that the general solution in that particular constant case is

$$G^{\boldsymbol{\nu},t}(\boldsymbol{x} - \boldsymbol{x}') = \frac{1}{2\pi|\boldsymbol{\Gamma}|^{1/2}} exp\left(-\frac{1}{2}\left(\boldsymbol{x} - \boldsymbol{x}'\right)^T \boldsymbol{\Gamma}^{-1}\left(\boldsymbol{x} - \boldsymbol{x}'\right)\right), \tag{8}$$

where

$$\boldsymbol{\Gamma} = 2t\boldsymbol{\nu}, \tag{9}$$

and $|\boldsymbol{\Gamma}|$ is the determinant of $\boldsymbol{\Gamma}$.

In this paper, the inverse of the $\boldsymbol{\Gamma}$ tensor will be expanded as

$$\boldsymbol{\Gamma}^{-1} = \begin{pmatrix} \frac{1}{L_x^2} & \frac{1}{L_{xy}} \\ \frac{1}{L_{xy}} & \frac{1}{L_y^2} \end{pmatrix},$$

Note that the real scalar $1/L_{xy}$ can be equal to zero, which corresponds to the particular case where $\boldsymbol{\Gamma}$ is diagonal, meaning that $\boldsymbol{\nu}$ is diagonal. The scale L_x (resp. L_y) corresponds to the one-dimensional differential length-scale along the direction x (resp. y) (Daley, 1991).

The normalization terms $\boldsymbol{\Lambda}$ is another important feature brought by the analytical solution Eq.(8). It appears that the kernel value for $\boldsymbol{x} = \boldsymbol{x}'$ is not 1 but $1/2\pi|\boldsymbol{\Gamma}|^{1/2}$. Thus, the normalization which will ensure that the standard deviation modelled to be 1, reads

$$\boldsymbol{\Lambda}^2 = 2\pi|\boldsymbol{\Gamma}|^{1/2}\mathbf{I}. \tag{10}$$

As expected, the resolution of the constant diffusion equation leads to Gaussian functions. However, in the non-constant diffusion equation, the *a priori* solution is no more Gaussian, but it is quasi-Gaussian with large scale geographical variations of the diffusion tensor.

2.4 Approximation of the local diffusion tensor and of the normalization in heterogeneous framework

In practice, the local diffusion tensor $\boldsymbol{\nu}(\boldsymbol{x})$ is not known. The dynamic aspects of the climatology of the geophysical flow can be considered to approximate the real diffusion tensor. But such an estimation does not take into account the real statistics of the background error. In particular, the time evolution of $\mathbf{B}$ is not represented, the heterogeneity of the observational network also influences the background error (Bouttier, 1994).

In this study, we are taking advantage of equations (9) and (10) to objectively approximate the local diffusion tensor $\boldsymbol{\nu}(\boldsymbol{x})$ and the local normalization $\boldsymbol{\Lambda}(\boldsymbol{x})$ for a given position $\boldsymbol{x}$.

The methodology is as follow: the estimation of the local matrix $\boldsymbol{\Gamma}^{-1}(\boldsymbol{x})$ leads to the local diffusion $\boldsymbol{\nu}(\boldsymbol{x}) = \boldsymbol{\Gamma}(\boldsymbol{x})/2$ and to the local normalization $\boldsymbol{\Lambda}^2(\boldsymbol{x}) = 2\pi|\boldsymbol{\Gamma}(\boldsymbol{x})|^{1/2}$. This methodology is appropriate for scales L relevant in meteorology and in oceanography for which the ratio $\gamma = a^2/L^2 \gg 1$, where a is the Earth radius (see appendix of Weaver and Courtier, 2001). In the remaining of the paper, t is set to 1.

The next section describes how to estimate $\boldsymbol{\Gamma}^{-1}(\boldsymbol{x})$ from a correlation function.

3 Estimation of the local diffusion tensor

The estimation of the local diffusion tensor $\boldsymbol{\nu}$ is deduced from the estimation of the local tensor $\boldsymbol{\Gamma}^{-1}$. The calculus of $\boldsymbol{\Gamma}^{-1}$ is achieved in two steps: the computation of the diagonal terms L_x and L_y is needed for the computation of the extra-diagonal terms L_{xy}. This is described in the following two sections.

3.1 Estimation of the diagonal components

The inverse of the diagonal terms of $\boldsymbol{\Gamma}^{-1}$ corresponds to the length-scales. These scales can be approximated with several low numerical cost formulae as described in Pannekoucke *et al.*(2008). Such length-scale approximations are well designed for various type of domain, like circle, plane, 2D or 3D-sphere. One of these formulae is based on the approximation of a correlation function ρ by a Gaussian.

For a specific direction, these approximations of the length-scales can be defined as follows. Let $\boldsymbol{\delta x}$ be the displacement in a direction $\boldsymbol{u} = \boldsymbol{\delta x}/|\boldsymbol{\delta x}|$ of the domain, the Gaussian approximation of the correlation function ρ leads to $\rho(\boldsymbol{\delta x}) \approx e^{-|\boldsymbol{\delta x}|^2/2L_{\boldsymbol{u}}^2}$. It follows that the Gaussian-based (Gb) length-scale is

$$L_{\boldsymbol{u}} = \frac{|\boldsymbol{\delta x}|}{\sqrt{-2\ \ln\ \rho(\boldsymbol{\delta x})}}, \tag{11}$$

with $|\boldsymbol{\delta x}|$ the magnitude of displacement.

In the particular case of a 1D domain, a left ($L_{-\delta x}$) length-scale and a right ($L_{+\delta x}$) length-scale can be defined. Thereafter, the left directional length-scale is designed by an superscript $^-$ and the right one by the superscript $^+$.

Thus, L_x and L_y, which correspond respectively to the zonal and to the meridional length-scales, can be estimated as $L_x = (L_x^+ + L_x^-)/2$ and $L_y = (L_y^+ + L_y^-)/2$. Now, the remaining non-diagonal terms have to be estimated.

3.2 Estimation of the non-diagonal term

The non-diagonal terms can be estimated from an approximation similar to the one used for the computation of the length-scales. In that case, the computation of the coefficient $1/L_{xy}$ is obtained from the approximation of the 2D correlation function ρ by a Gaussian function

$$\rho(\boldsymbol{x}) \equiv exp\left(-\frac{1}{2}\boldsymbol{x}^T\boldsymbol{\Gamma}^{-1}\boldsymbol{x}\right).$$

Then, for a given direction and distance $\boldsymbol{\delta x} = (\delta x, \delta y)$ with $\delta x \neq 0$ and $\delta y \neq 0$, this term is estimated as

$$\frac{1}{L_{xy}} = \frac{-1}{2\delta x \delta y}\left[2\,ln\,\rho\left(\boldsymbol{\delta x}\right) + \frac{\delta x^2}{L_x^2} + \frac{\delta y^2}{L_y^2}\right]. \qquad (12)$$

Finally, the local tensor $\boldsymbol{\Gamma}^{-1}(\boldsymbol{x})$ can be computed from its diagonal and non-diagonal terms previously described. This leads to the local diffusion tensor

$$\boldsymbol{\nu}(\boldsymbol{x}) = \boldsymbol{\Gamma}(\boldsymbol{x})/2,$$

and to the local normalization

$$\boldsymbol{\Lambda}^2(\boldsymbol{x}) = 2\pi|\boldsymbol{\Gamma}(\boldsymbol{x})|^{1/2}.$$

4 Illustration on the circle

The approximation of the diffusion coefficient and the normalization are tested in a one dimensional framework. In this case, the local tensor $\boldsymbol{\Gamma}^{-1}(\boldsymbol{x})$ is restricted to a scalar $1/L_x^2$.

The example is constructed as follows. In a first step, a non-trivial length-scale field is build. Then, this field is used to construct a heterogeneous covariance matrix with the non-homogeneous diffusion equation, but without the normalization. The resulting length-scale is compared with the initial length-scale. Finally, the normalization computed from the initial length-scale is compared with the found numerical normalization.

The domain is a circle of radius $a = 6480\,km$, which corresponds to the Earth great circle. The circle is divided into $n_g = 241$ equally-spaced grid-points, associated to a truncation $T = 120$.

4.1 Generation of a length-scale field

The length-scale field is generated as the sum of a mean length-scale value L_h plus a length-scale perturbation δL, sampled from a given energy-spectrum. The perturbation δL is designed so that its standard deviation, for a given position, is $\sigma = 0.2\,L_h$. In the following, the energy spectrum of the perturbation is set as

$$\begin{cases} E(n) = 0 & \text{for } n = 0, \\ E(n) = \lambda\, n^p (1 + \frac{n}{n_c})^{q-p} & \text{elsewhere,} \end{cases}$$

where $p > 0$ corresponds to the positive energy slope in the large scales, $q < 0$ corresponds to the negative energy slope in the small scales, n_c is a cut off wave number and λ a normalization to ensure that the total energy is $E_{tot} = \sigma^2$. A perturbation δL is computed as a Gaussian random realization whose energy spectrum is $E(n)$ (under an ensemble average).

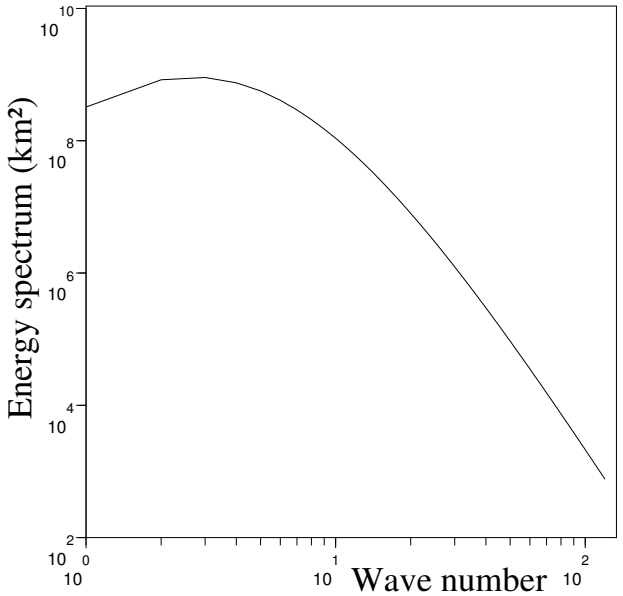

Figure 1. Example of energy spectrum for the perturbation δL used to generate the length-scale field, for the particular set of parameters $L_h = 350\,km$, $p = 4$, $n_c = 4$ and $q = -6$ (see text for details).

Figure 1 illustrates such an energy spectrum for the set $L_h = 350\,km$, $p = 4$, $n_c = 4$ and $q = -6$. A particular sample associated to this spectrum and to the L_h value is represented at the top panel of Figure 2 (solid line). This length-scale field is denoted L_{th}.

4.2 The modelled heterogeneous covariance matrix

From the length-scale field $L_{th}(x)$, the field of diffusion coefficients, corresponding to Eq. (9), is computed as $\nu(x) = \frac{L_{th}(x)^2}{2\tau}$, with $\tau = 1$. The diffusion operator $\mathbf{L}^{1/2}$ corresponds to the propagator of the diffusion equation

$$\partial_t u = \partial_x\left(\nu \partial_x u\right), \qquad (13)$$

from $t = 0$ to $t = 1/2$. In the numerical experiments, $\mathbf{L}^{1/2}$ is explicitly computed as the exponential of the discrete problem (see appendix B for details).

In the regularly discretized circular framework, the metric tensor is featured by the diagonal $\mathbf{W}^{-1} = n_g/2\pi a\,\mathbf{I}$ (see appendix C).

The modelled covariance matrix based on the diffusion operator is then given by Eq. (5). The matrix built with the square-root of the diagonal elements of $\tilde{\mathbf{C}}$ corresponds to the inverse of the normalization $\boldsymbol{\Lambda}$. Thus the correlation matrix modelled with the diffusion operator is $\mathbf{C} = \boldsymbol{\Lambda}\mathbf{L}^{1/2}\mathbf{W}^{-1}\mathbf{L}^{T/2}\boldsymbol{\Lambda}^T$.

4.3 Diagnosis of the length-scale and validation

The length-scale of the modelled correlation matrix, in the circle framework, can be diagnosed either with Daley

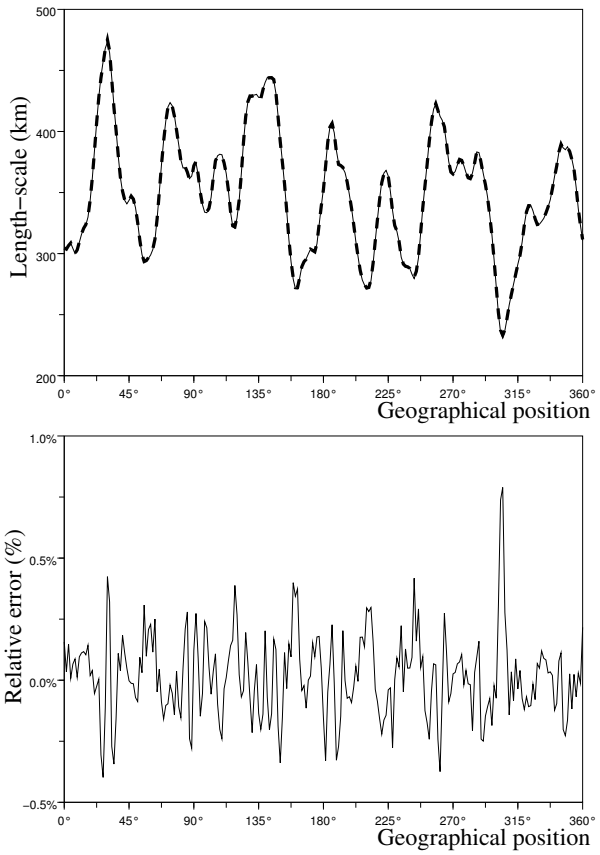

Figure 2. Top panel: A particular length-scale field L_{th} generated from the spectrum shown in Fig.1 (solid line) and the mean Gb length-scale diagnosed L_{diag} from the covariance modelling using the heterogeneous diffusion operator on the circle, with diffusion coefficient computed from L_{th} (bold dashed line). Bottom panel: Relative error $e_{\%}$ (in percent) between L_{th} and L_{diag}.

formula of the length-scale or with the mean Gb length-scale computed as $L_{diag} = \left(L_x^+ + L_x^-\right)/2$. Both types of length-scale estimation lead to similar length-scale values (not shown here).

The diagnosed length-scales L_{diag} are represented at the bottom panel of Figure 2 (bold dashed line). It appears that L_{th} and L_{diag} are very close. The differences between the two fields is measured by their relative error $e(x) = 100\left(L_{th} - L_{diag}\right)/L_{th}$. The usual error is less than 0.5%. The error field feature is slightly related to the length-scale field: the error is maximal (in absolute value) for rapid variation of the length-scale field *e.g.* in the vicinity of $90\,^\circ$; or for extreme length-scale values *e.g.* for the large peaks near $33\,^\circ$ or $300\,^\circ$.

Another interesting diagnosis is the difference between the energy spectrum of the two length-scale fields L_{th} and L_{diag} (figure 3). The energy spectrum of L_{diag} (dotted line) is close to the energy spectrum of L_{th} (solid line) for a wide range of wave numbers. Again, this illustrates the accuracy of the method.

Finally, this example illustrates the ability of the matrix modelled with the diffusion operator to represent the initial length-scale variations. This proves that inverting the relation between the diffusion coefficient and the length-scale (exact in homogeneous cases, and taken as an

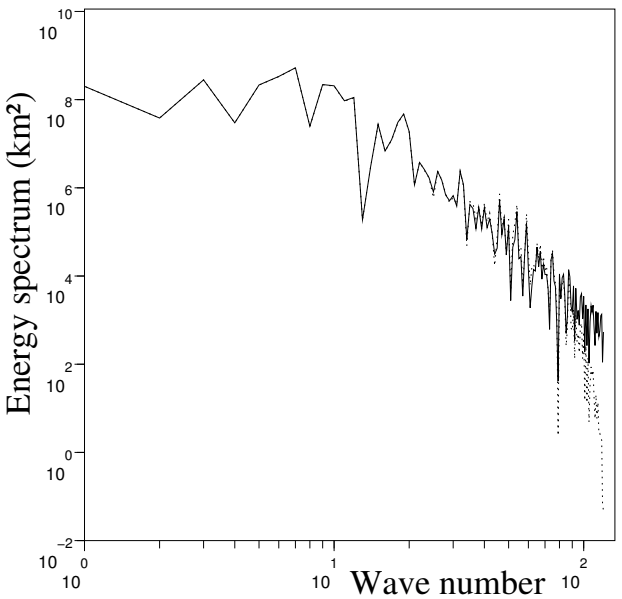

Figure 3. Energy spectrum of initial length-scale field L_{th} (solid line) compared to the diagnosed length-scale field L_{diag} of modelled covariance matrix with the diffusion operator (dotted line).

approximation in non-homogeneous cases) leads to a correlation matrix whose length-scale L_{diag} corresponds to the initial L_{th}.

In the following paragraph, we are going to show how the normalization $\mathbf{\Lambda}$ can be approximate from the length-scale.

4.4 Validation of the normalization based on the length-scale

As reminded in section 2.3, in the case of a constant diffusion tensor, the normalization is related to the $\mathbf{\Gamma}^{-1}$ tensor through Eq. (10). But in the particular case of a one dimensional framework, the homogeneous normalization is simply related to the length-scale L by $\mathbf{\Lambda}^2 = \sqrt{2\pi}L\mathbf{I}$. In a non-homogeneous one dimensional case, $\mathbf{\Lambda}^2$ is approximated by the diagonal built with the normalization $\sqrt{2\pi}L(x)$ related to a given position x.

Figure 4 represents the product of the diagonal of $\tilde{\mathbf{C}}$ (defined by Eq. 5) by the normalization $\sqrt{2\pi}L_{th}(x)$. If this normalization was the true one, then the resulting product should be everywhere equal to 1. Here, it appears that the product is equal to 1 with an accuracy of less than 5%.

In conclusion, the approximation of the normalization, based on the local length-scale value, is an accurate estimation of the true normalization. Nevertheless, another strategy to estimate such a normalization can also be used: the approximation based on the "Parametrix method" (Purser *et al.*, 2007).

4.5 Sensitivity to the energy spectrum shape

Various experiments have been caried out to study the sensitivity of the length-scale and of the normalization estimation (computed as described in previous sections) to the shape of the initial length-scale energy spectrum. These

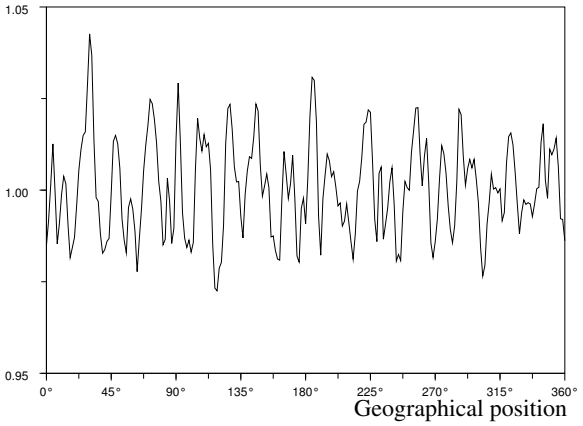

Figure 4. Variances resulting by applying the normalization factor deduced from the length-scale.

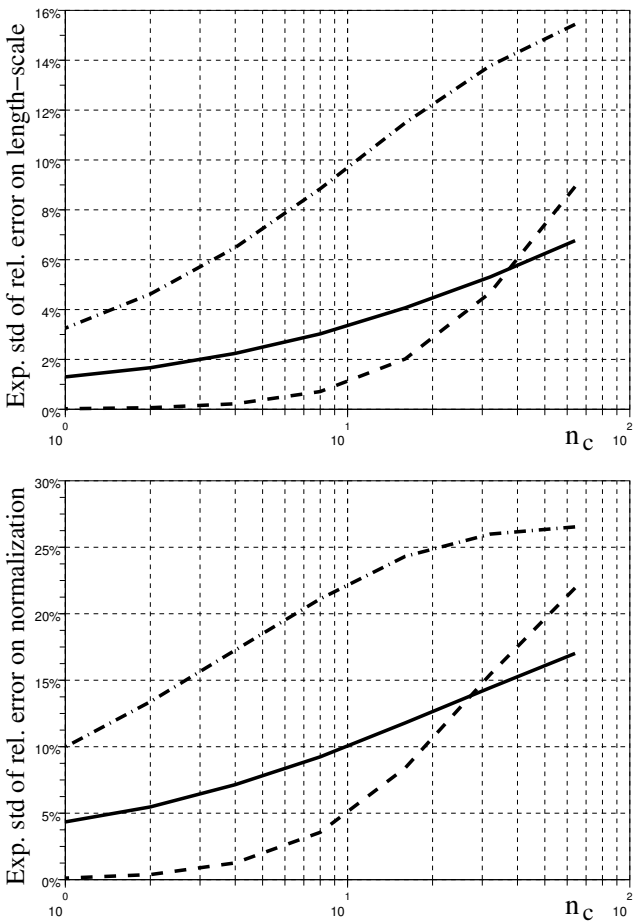

Figure 5. Expected standard deviation of the relative error on the length-scale (top) and on the normalization (bottom) for $(p = 0, q = -2)$ (solid line), $(p = 4, q = -6)$ (dashed line) and $(p = 8, q = -2)$ (dash-dotted line) versus the cut-off wave number $n_c \in [1, 64]$.

experiments were performed by changing the parameters (p, q, n_c) of the spectrum defined in section 4.1.

For each set (p, q, n_c), $Ns = 400$ independent length-scale fields have been generated. The construction of such a length-scale field has been designed so that the shortest length-scale value is larger than the grid resolution.

Figure 5 represents the standard deviation of the relative error for the length-scale and for the normalization, computed from the generated length-scale fields. These experiments have been realized for three different shapes of length-scale spectrum, with a cut-off number varying from 1 to 64.

The first shape, defined by $(p = 0, q = -2)$ (solid line), corresponds to a white spectrum until the cut off, from which the energy decreases slowly. This shape is associated to a red spectrum (RS). The second shape is defined by $(p = 4, q = -6)$ (dashed line). In this case, the spectrum increases rapidly until the cut-off number; a fast decrease then follows the end of the spectrum. The length-scale field is featured by an average scale closed to the cut-off number. This type of signal is denoted by PS (spectrum with peak). The third shape is associated to a blue spectrum defined by $(p = 8, q = -2)$ (dashed-dotted line). All these shapes lead to a red spectrum for small n_c and to a blue spectrum for large n_c. This last shape is associated to a blue spectrum (BS).

The bulk variations correspond to an error increasing with the cut-off wave number. It appears that the discrepancy is larger when the spectrum is blue. For $n_c \geq 30$, the error affecting the length-scale is larger than $\approx 6\%$. For the normalization, the error is larger than $\approx 15\%$. For red spectrums (corresponding to small n_c), the error on the length-scale is less than 10% for $n_c \leq 10$, while the error on the normalization is less than 25%.

This type of variations are also encountered among the various shapes: for small large scale ($n_c \leq 20$), the error for PS is lower than the error for RS, itself lower than the error for BS.

It results that the estimation of the diffusion coefficients and of the normalization from the length-scale field is accurate in a case of smooth length-scale fields, *i.e.* a length-scale field dominated by large scale components. This can appear as a limitation of this approach. However, in practice, the length-scale is estimated from finite ensemble. This estimation is known to be affected by the sampling noise, featured by spurious small scales contribution (Pannekoucke *et al.*, 2008). A possible way to reduce the sampling noise is to filter the small scale of the length-scale field, *e.g.* with a spatial filtering (Berre *et al.*, 2007).

5 Illustration on the sphere: MOCAGE-PALM background error covariances

A one dimensional study was presented in the previous section. To explore the strengths of the method described above, a three dimensional study was performed using the MOCAGE-PALM chemical-transport assimilation system. The corresponding $\mathbf{\Gamma}$ tensor and the normalization were computed for tropospheric and stratospheric ozone content. Based on these results, a heterogeneous covariance matrix was modelled and tested.

5.1 MOCAGE-PALM assimilation system

The assimilation system used in this study is derived from the MOCAGE-PALM system developed jointly by CERFACS and Météo-France in the framework of the FP5 European project ASSET (Lahoz *et al.*, 2007). The assimilation algorithm is based on a 3D-VAR algorithm, in the FGAT (first guess at appropriate time) variant (Fisher and Anderson, 2001). The system is based on the Météo-France comprehensive three-dimensional chemistry transport model (CTM) MOCAGE and on the CERFACS PALM software (Buis *et al.*, 2006). The CTM MOCAGE covers the planetary boundary layer, the free troposphere and the stratosphere. It provides a number of optional configurations with varying domain geometries and resolutions, as well as chemical and physical parameterization packages. MOCAGE is currently used for several applications, such as in chemical weather forecasting (Dufour *et al.*, 2004), chemistry–climate interactions (Teyssèdre *et al.*, 2007) and data assimilation (Pradier *et al.*, 2006; Massart *et al.*, 2005a and 2005b). The first version of the MOCAGE-PALM assimilation system, as it was originally implemented for the ASSET project, provided good quality ozone fields compared with ozonesondes and UARS/HALOE measurements with errors of the same order as those supplied by several other assimilation systems (Geer *et al.*, 2006). In order to improve the assimilation system, several changes have been recently made on the model resolution and on the background error characterization (Massart *et al.*, 2007). Thus, in this study, the domain geometry and resolution were a global $2\,^\circ \times 2\,^\circ$ horizontal grid and a 60 level hybrid (σ, P) vertical discretization from the surface up to 0.1hPa. The meteorological forcing fields were provided by the operational European Centre for Medium-Range Weather Forecasts (ECMWF) numerical weather prediction model. We have also adopted the linear ozone parameterization developed by Cariolle and Teyssèdre (2007) in its latest version. In fact, the background error covariance matrix of the MOCAGE-PALM assimilation system is divided into an horizontal and a vertical operators. The vertical correlation is modelled using a Gaussian formulation in terms of the logarithm of the pressure. The horizontal correlation is modelled using a two dimensional diffusion equation (Weaver and Courtier, 2001) with a homogeneous length-scale of 4 degrees (that corresponds to a distance of approximately 445 km).

5.2 The used ensemble data set

The estimation of the background error correlations using an ensemble of assimilation has proved to be an efficient method (Belo Pereira and Berre, 2006). The ensemble used in this study is based on ten sets of perturbed observations derived from the Envisat/MIPAS reference data set of ASSET (version 4.61 delivered by the German Processing and Archiving Center, D-PAC, of the European Space Agency). The Michelson Interferometer for Passive Atmospheric Sounding (MIPAS) instrument gives 24h a day ozone profiles with a very good global coverage, extending vertically from above the top of our model downward to around 300 hPa, with a 3km to 8km resolution (Raspollini *et al.*2006).

The ten perturbed observation sets are obtained by adding to the July 2003 MIPAS data, random perturbations which are drawn from a Gaussian distribution based on the covariance matrix specified by the D-PAC. They permit to realize an ensemble of 10 members of 3-h ozone forecasts from the MOCAGE-PALM assimilation system. In order to eliminate the transient period, and to ensure a sufficient dispersion of the ensemble, only the last twenty days of July 2003 are used to compute the statistics. The resulting ensemble over this period counts 1600 (20 days $\times$ 8 forecasts of 3-h per day $\times$ 10 forecasts members) members of forecast errors.

5.3 Estimation of the local diffusion tensor for the ozone and its features: length-scales and anisotropy vectors

As explained in Belo Pereira and Berre (2006), the previously described ensemble of perturbed forecast is used to estimate the background error correlation in the grid-points. Spatial correlations of the background error are then computed over the whole MOCAGE spatial domain. Some of these functions are illustrated in the left panel of figure 6. It firstly shows that the correlation functions differ from one point to another with quite a similarity over parallel lines. Moreover, the correlation functions seem to be more isotropic over the South Polar region than to the high latitude region. These primary diagnoses will be confirmed by the evaluation of L_x, L_y and L_{xy} from the spatial correlations as described in section 3.

Since the estimation of the length-scales is affected by the sampling noise in the small scales (Pannekoucke *et al.*, 2008), the noisy contribution can be filtered by a spatial average (Berre *et al.*, 2007). Thus, the diagnosed length-scales L_x and L_y (not L_{xy}) are filtered by a convolution with a radial function. This has been numerically achieved by using an isotropic diffusion function with an associated length of $450\,km$. The effect in terms of energy spectrum is illustrated Figure 8 for the raw estimation (solid line) and the filtered L_x length-scale. As expected, the filter dampes a large part of the small scales present in the raw estimation. Similar spectrum is obtained for L_y. Note that the filter used in this paper is not optimal as described by Berre *et al.*(2007).

The result for L_x (resp. L_y) is represented on the left (resp. right) top panel of Figure 7, for the ozone at the pressure level 10 hPa. This pressure level was selected to illustrate the length-scales because it offers a good representation of the phenomena observed within the stratosphere (where most of the observations used to produce the forecast members are located). The top panel of Figure 7 first illustrates the variation of both the L_x and L_y length-scales over the globe.

Probably due to the earth dynamics characterized by a transport mainly directed along latitude lines, the L_x length-scale is mostly greater than the L_y one. The L_x

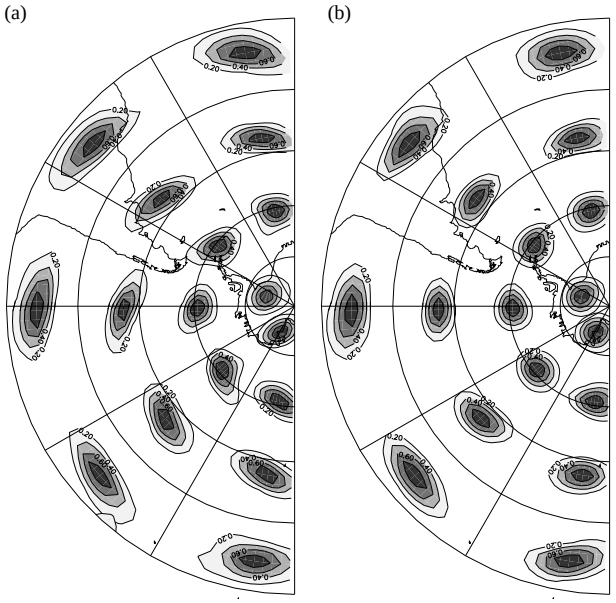

Figure 6. Examples of correlation functions of the background error viewed from the South Pole: (a) diagnosed from the ensemble of perturbed forecast and (b) from the correlation model diagnosed using a randomization method with 6400 samples. The contour lines correspond to values within the range 0.2 – 1, with an 0.2 increment. Parallel lines are plotted every 15 degrees from $90\,^{\circ}S$ to $15\,^{\circ}S$.

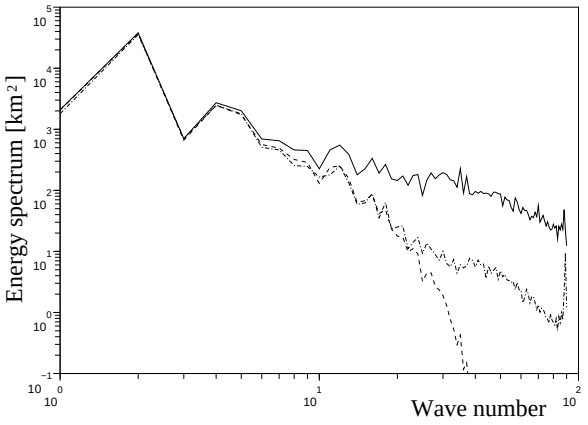

Figure 8. Energy spectrums of the MOCAGE-PALM L_x length-scale estimated by the ensemble of forecast (solid-line), of this length-scale filtered with an isotropic diffusion function (dashed line) and of the length-scale from the formulation diagnosed using a randomization method (dash-dotted line).

length-scale is quite constant over the parallels but has significant variations along the meridians with decreasing values from around $600\,km$ at the equator to $300\,km$ at the poles. The L_y length-scale has a more notable variation along the parallels with values between around 200 and $350\,km$.

The anisotropy vectors diagnosis is considered similarly to Belo Pereira and Berre (2006). The local anisotropy vector corresponds to the leading principal axe of the diffusion tensor $\boldsymbol{\nu}(\boldsymbol{x})$. The norm of the anisotropy vector corresponds to the oblateness $1-\lambda_2/\lambda_1$ with λ_1 and λ_2 respectively the largest and the smallest eigenvalue of $\boldsymbol{\nu}(\boldsymbol{x})$. Both the norm of the anisotropy vector and the vector itself are represented in Figure 9. As noticed for L_x, the anisotropy intensity is essentially a zonal phenomenon with high values in the equatorial and tropical regions and with lower values over the poles. However, one can observe the development of a wave in the South Polar Region with local maximal values. This wave is associated with an anisotropy pointing in the South-Nord direction while the main direction is West-East elsewhere.

5.4 Modelled heterogenous covariances

The computation of the local tensor $\boldsymbol{\Gamma}^{-1}(\boldsymbol{x})$ from our estimation of L_x, L_y and L_{xy} at each position $\boldsymbol{x}$, as described in section 3, allows us to compute the local diffusion tensor $\boldsymbol{\nu}(\boldsymbol{x})$ using Eq. (9).

The heterogeneous covariance matrix $\tilde{\mathbf{C}}$ of Eq. (5) based on the diffusion operator and on the local diffusion tensor $\boldsymbol{\nu}(\boldsymbol{x})$ has been computed over the sphere. The propagator $\mathbf{L}^{1/2}$ corresponds to the integration of Eq. (5). The time discretization corresponds to a forward Euler scheme. The differentials operators have been computed in the spectral space. A classical $2/3$ filter has been used to eliminate the aliasing effect resulting from the computation of the product between the diffusion and the gradient (Boyd, 2001). The time step has been chosen in order to ensure the Courant Friedriech Levy condition of stability.

In the particular case of a triangular truncation, the operator $\mathbf{W}$ is featured by the diagonal matrix

$$\mathbf{W}^{-1} = \frac{(T+1)^2}{4\pi a^2}\,\mathbf{I},$$

where T is the truncation and a the radius of the Earth (see appendix C).

5.5 Diagnosis of the length-scales and the anisotropy vectors

The diagnosis of the modelled heterogeneous background correlation matrix $\tilde{\mathbf{C}}$ can be achieved with an exact computation of each modelled correlation function (as in the 1D framework) or with a randomization method (Fisher and Courtier, 1995; Weaver and Ricci, 2003). In this part of the work, the second method has been considered. Thus an ensemble of background error has been generated according to $\boldsymbol{\varepsilon}^b = \tilde{\mathbf{C}}^{1/2}\boldsymbol{\zeta}$ where $\boldsymbol{\zeta}$ is a realization of a centred Gaussian perturbation with the identity matrix as its covariance matrix. Some correlation functions from the randomization method are illustrated in the right panel of figure 6. The comparison of the two panels shows that the modelled correlation functions well reproduce the anisotropy and the heterogeneity. Likewise, in the neighbourly of the core of each correlation functions, the model well represents the amplitude of the correlations in the two dimensions.

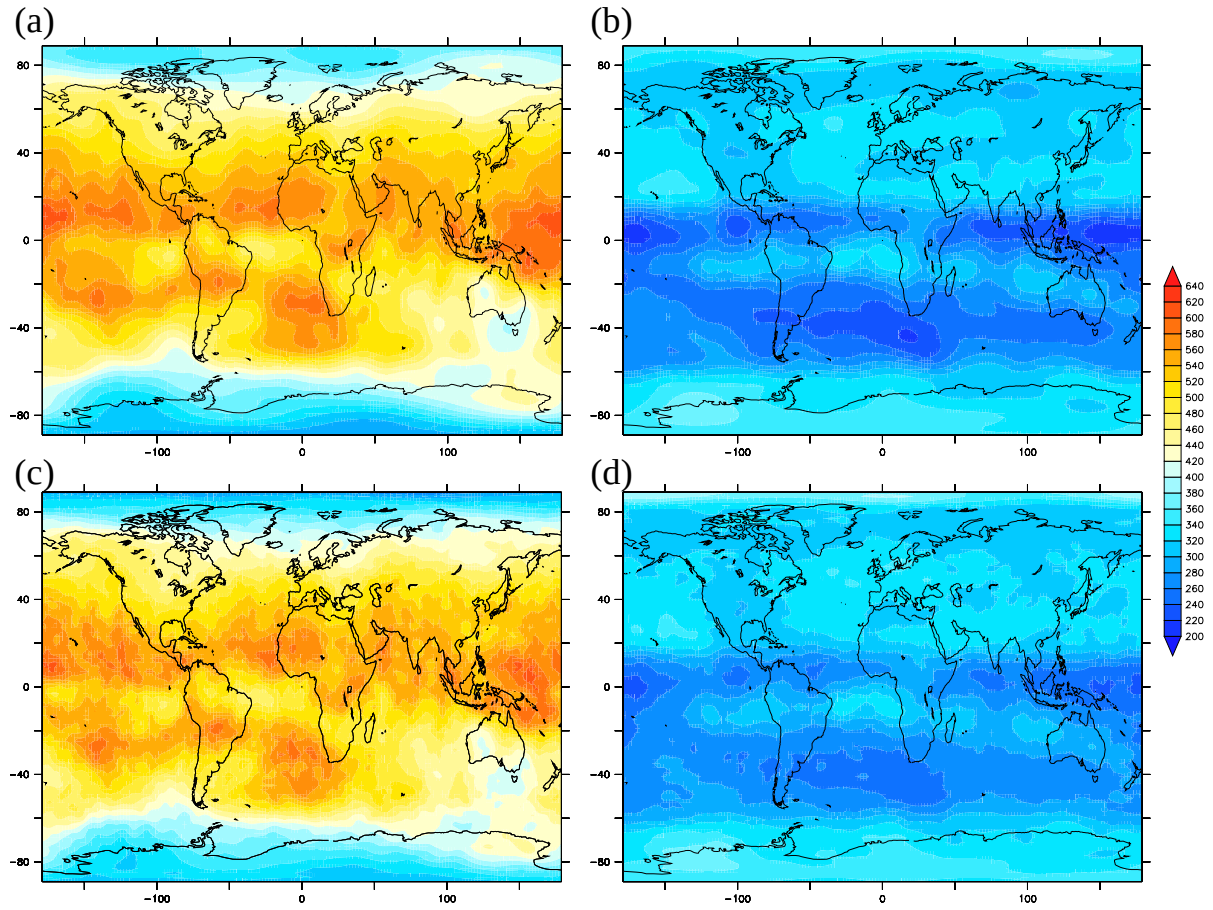

Figure 7. Representation of the raw length-scale estimation over the last twenty days of July 2003 at 10hPa: zonal (a) and meridional (b) length-scale. These field are used to approximate the diffusion coefficients associated to such a heterogeneity. The zonal (c) and meridional (d) length-scale of the resulted formulation are diagnosed using a randomization method with 6400 samples.

Similarly to the estimation of the raw length-scales, the length-scales of $\tilde{\mathbf{C}}$ have been estimated from the generated ensemble by using the Gaussian-based approximation defined by Eqs. (11) and (12). In the experiment, a large ensemble of 6400 members has been generated so that the sampling noise affecting the estimation is less than 2%.

The energy spectrum of the zonal length-scaled iagnosed from the generated ensemble is represented Figure 8 (dash-dotted line). In comparison with the filtered length-scale (dashed line), both spectrums are very close until total wave numbers around 20. For higher total wave numbers, the sampling noise adds energy in the small scales. Thus, the the spectrum of the length-scale from the formulation diagnosed using the randomization method decreases at a lower rate than the one from the filtered length-scale. The zonal and the meridional length-scales diagnosed from the generated ensemble are represented in the physical space Figures 7-(c) and (d). The diagnosed length-scale fields are very close to the initial raw estimation (figures 7-(a) and (b)). In details, the correlation matrix $\tilde{\mathbf{C}}$ underestimates the zonal length-scales with an average relative error of 0.4%. These errors are maximal over the poles with values of 8% at the South and -15% at the North. This has to be linked with lower length-scale values over the poles and to the difficulty of the method to catch small values (compared to the size of the mesh). Concerning the meridional length-scales, the diagnosis reveals that the correlation matrix $\tilde{\mathbf{C}}$ overestimates the length-scales with an average relative error of 2.2%. This error is higher than the one of the zonal length-scales, probably beacause for the reason that the length-scales along the meridians have lower values than those along the parallels. Moreover, as for the zonal length-scales, the relative error between the diagnosed and the raw meridional length-scales reaches its maximum when the length-scale values are the lowest. Then, the relative error is about 8% in two latitude bands, one, just above the equator, and the other, below the equator.

Figures 9-(c) and (d) represent respectively the norm and the direction of the anisotropy diagnosed from the same ensemble as the one use for the length-scales. It appears that the diagnosed anisotropy is close to the initial raw estimation with an average underestimation of 2.5%. In this case, the error is maximal with values around 80% where the anisotropy has important gradient in the South pole region. As expected from the 1D framework results, rapid geographical variations, associated to small

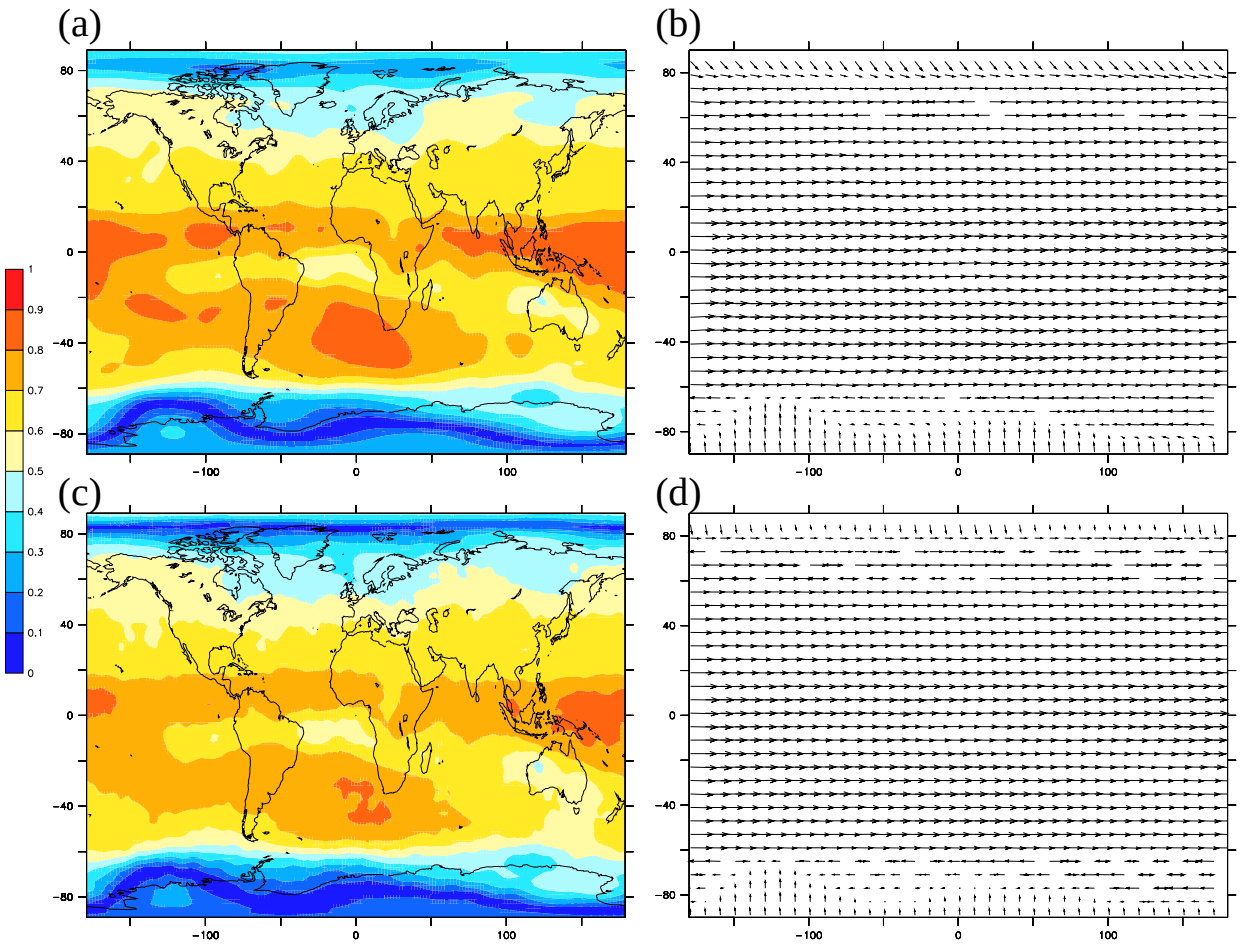

Figure 9. Representation of the anisotropy estimation (resp. modelled with diffusion operator) over the last twenty days of July 2003 at 10hPa: norm (a) and direction (b) of the anisotropy vector (resp. (c) and (d)).

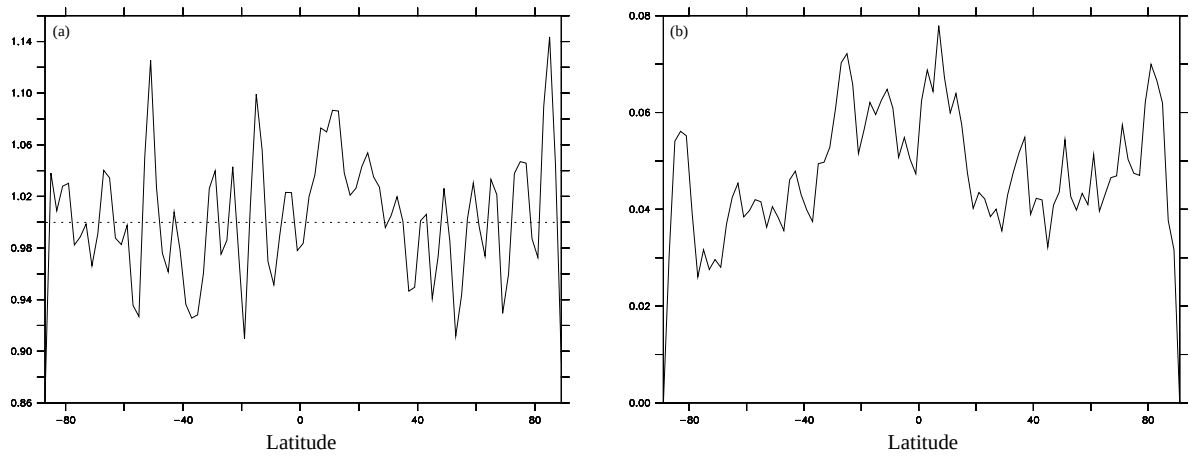

Figure 10. Validation of the factor deduced from the local diffusion tensor using the randomization method with 6400 samples: (a) variance along longitude 0° and (b) standard deviation of the variance.

scale contributions in the initial length-scale field, are attenuated.

5.6 Validation of the normalization based on the length-scales

As the validation of the normalization in the case of the circle, section 4.4, the normalization based on the length-scales is computed here using Eq. (10). With the same

randomization method used previously for the length-scales, the diagonal of $\tilde{\mathbf{C}}$ (defined by Eq. 5) is estimated with a 6400 samples. The product of this diagonal by the normalization $2\pi|\mathbf{\Gamma}(\boldsymbol{x})|^{1/2}$ is computed and has to be compared to the unity. This porduct is illustrated by Figure 10 along the longitude 0° and its standard deviation as a function of latitude. As shown, the predetermined normalization is a very good approximation of the true one (with an error less than 0.3%) more or less 5%.

As a result, the normalization based on the local diffusion tensor offers an accurate approximation of the true normalization. Note that the accuracy is reached in a randomization process since our ensemble size is greater than the one needed (1000) as discussed by Weaver and Ricci (2003).

6 Conclusion

In this paper, we have described the estimation of the local diffusion tensor and the local normalization appearing in the modelling of a heterogeneous background error covariance matrix based on the diffusion equation. This estimation is based on the relationship between the local diffusion tensor and the local length-scales. It has been shown that the use of the estimated diffusion tensor leads to a formulation of the background error covariance matrix that reproduces the initial feature of the background errors. The link between the diffusion tensor and the length-scales is well-known in the case of a constant diffusion tensor. This study shows that this link is still valid in an heterogeneous case when the geographical variations are relatively smooth.

This has been firstly illustrated for a simple one dimensional framework. The local diffusion coefficients have been estimated from a given length-scale field that varies along the domain. These coefficients have then been used to model a heterogeneous covariance matrix based on the diffusion operator. It has been shown that the diagnosed length-scale field from this modelled matrix corresponds to the initial length-scale field. This correspondence is particularly accurate when the spectrum of the initial length-scale field is red. Similar conclusions have been obtained for the normalization field.

Then, the method has been illustrated for a specific altitude level of a three-dimensional global chemistry transport model. An ensemble of model forecasts has permitted to estimate the length-scales and the anisotropy vector of the ozone forecast errors. The local diffusion tensor was deduced from these estimations. A two dimensional heterogeneous background error covariance matrix was modelled, based of this local diffusion tensor. As for the one dimensional study, the diagnosed length-scale fields from this modelled matrix have been shown to correspond to the initial length-scales with a low average relative error. The diagnosis has also revealed that it is more difficult to model efficently the anisotropy in this specific case where spatial variations are important. Concerning the normalization, the values given by the length-scales tend to be close to the expected ones. This proposed normalization is especially interesting since there is no other low cost way to compute it in the particular heterogeneous case.

This two dimensional application was the first step towards the appliance to the whole three dimensional domain of our chemistry-transport model. Nevertheless, the three dimensional error covariance operator will be separated on each grid point into a two dimensional horizontal one and a one dimensional vertical one. The full evaluation of this methodology has now to be carried on by assessing the impact of first diagnosed and then modelled local covariance on the analysis and finally on the forecast. Moreover, this methodology could also be applied to full scale atmospheric and oceanic forecast system. As there is no computational cost associated to the normalization, the main cost is therefore attributed to the computation of the members (and a sufficient number of members is required to have significant statistics when the local length-scale are computed). Furthermore, an efficient diffusion solver is also an important requirement.

Acknowledgement

The authors would like to thanks Anthony Weaver and Nicolas Daget for fruitful discussions, Laura Pebernet for her contribution and Jean-Antoine Maziejewski.

A Analytical resolution in the plane of the constant diffusion equation

The expression of the kernel $G^{\boldsymbol{\nu},t}$, solution of 2D equation Eq.(2) is sought. The diffusion tensor $\boldsymbol{\nu}$ of Eq.(7) is assumed to be symmetric, it follows that there exists an orthonormal basis of the plane where the matrix of the representation of $\boldsymbol{\nu}$ is diagonal *i.e.* $\nu_{x,y} = 0$. Thus, without lost of generality, $\nu_{x,y}$ is thereafter assumed to be null.

Let $G^{\nu_x,t}(x)$ and $G^{\nu_y,t}(y)$ be solutions of, respectively,

$$\begin{aligned} \partial_t G^{\nu_x,t} &= \nu_x \partial^2_{xx} G^{\nu_x,t}, \\ \partial_t G^{\nu_y,t} &= \nu_y \partial^2_{yy} G^{\nu_y,t}. \end{aligned}$$

Noticing that $G^{\boldsymbol{\nu},t}(x,y) = G^{\nu_x,t}(x) G^{\nu_y,t}(y)$ is a solution of Eq. (5), the 2D solution can be constructed as the product of 1D solutions. One can verify that a solution of the constant 1D diffusion equation on the real line, with a Dirac as initial condition, is

$$\begin{aligned} G^{\nu_x,t}(x) &= e^{-x^2/4\nu_x t}/\sqrt{4\pi\nu_x t}, \\ G^{\nu_y,t}(y) &= e^{-y^2/4\nu_y t}/\sqrt{4\pi\nu_y t}. \end{aligned}$$

The solution of the 2D problem is thus

$$G^{\boldsymbol{\nu},t}(\boldsymbol{x}) = \frac{1}{2\pi|\mathbf{\Gamma}|^{1/2}} exp\left(-\frac{1}{2}\boldsymbol{x}^T\mathbf{\Gamma}^{-1}\boldsymbol{x}\right), \qquad (14)$$

where $\mathbf{\Gamma} = 2t\boldsymbol{\nu}$ and $|\mathbf{\Gamma}|$ is the determinant of $\mathbf{\Gamma}$. The invariance by translation due to $\boldsymbol{\nu}-$constant assumption involve, $G^{\boldsymbol{\nu},t}(\boldsymbol{x},\boldsymbol{x}') = G^{\boldsymbol{\nu},t}(\boldsymbol{x}-\boldsymbol{x}')$.

B Resolution of the heterogeneous diffusion equation over the circle

The aim of this section is to evaluate $\mathbf{L}^{1/2}$ in Eq. (5) in the particular case of a one dimensional circle framework. It can be noted that $\mathbf{L}^{1/2}$ corresponds to the propagator from $t = 0$ to $t = 1/2$ associated to the equation (11). The domain represents the great-Earth circle that is regularly discretized with n points corresponding to the truncation T so that $n = 2T + 1$.

For the vector $\boldsymbol{u}$ corresponding to the discretized field over the circle, the discrete diffusion equation can be written

$$\frac{d\boldsymbol{u}}{dt} = \mathbf{S}^{-1}\mathbf{D}\,\mathbf{K}\,\mathbf{D}\mathbf{S}\boldsymbol{u}, \tag{15}$$

where $\mathbf{S}$ is the matrix that corresponds to the discrete Fourier transform, $\mathbf{D}$ is the diagonal matrix corresponding to the discrete version of the differential operator ∂_x, and $\mathbf{K}$ is the matrix representing the convolution in the spectral space that corresponds to the grid-point product with the field ν.

If in the spectral space, the spectral coefficients are ordered as $(u_p)_{p\in[-T,T]}$, then $\mathbf{D} = Diag\left\{(\imath p)_{p\in[-T,T]}\right\}$, and $\mathbf{K}$ is the $n \times n$ banded Hermitian Toeplitz matrix where the first column is the n-list $(\nu_0, \nu_1, \cdots, \nu_T, 0, \cdots, 0)$ and the first row is the n-list $(\nu_0, \nu_{-1}, \cdots, \nu_{-T}, 0, \cdots, 0)$, where $(\nu_p)_{p\in[-T,T]}$ is the spectrum of the discretized diffusion field. Note that the matrix $\mathbf{S}^{-1}\mathbf{D}\,\mathbf{K}\,\mathbf{DS}$ is Hermitian.

The solution of linear differential equation (15) is $u(t) = e^{t\mathbf{S}^{-1}\mathbf{D}\,\mathbf{K}\,\mathbf{DS}}u(0)$. Thus, $\mathbf{L}^{1/2} = e^{\frac{1}{2}\mathbf{S}^{-1}\mathbf{D}\,\mathbf{K}\,\mathbf{DS}}$. In numerical experiments, this exponential is explicitly computed.

C Expression of the metric on the isotropic circle and sphere

According to section 2.2, the normalization is defined from the distribution of Dirac δ. This distribution depends on the measure and the vectorial space where the solution is sought.

In the case of the $a-$radius homogeneous circle $\mathcal{C}$ associated to the truncation T and where the measure is defined as $d\omega = dx/2\pi a$. The aim is to find the expression of δ_l corresponding to the Dirac distribution at point l. Due to the translation invariance property, this distribution does not depend on the point of the circle. Thus, without lost of generality, only the distribution δ at point $x = 0$ is sought. δ is included in the subspace $\mathcal{F}_T$ span by exponential functions $e_p(x) = e^{ipx/a}$ for $p \in [-T, T]$. The exponential functions form an orthogonal basis of this subspace. Thus, the coordinate of δ along $e^{ipx/a}$ corresponds to the Fourier coefficient

$$\langle e_p|\delta\rangle = \frac{1}{2\pi a}\int_{\mathcal{C}} e^{-ipx/a}\delta dx.$$

By definition of the Dirac distribution

$$\frac{1}{2\pi a}\int_{\mathcal{C}} e^{-ipx/a}\delta dx = \frac{e^{-ip0/a}}{2\pi a}$$

that is equal to $1/2\pi a$. Thus, $\delta(x) = 1/2\pi a \sum_{p=-T}^{T} e_p(x)$ also equal to

$$\delta(x) = \frac{1}{2\pi a}\left(1 + \sum_{p=1}^{T} 2\cos\,\frac{px}{a}\right).$$

Moreover, the projection of $\chi_{\mathbf{0}}$ on $\mathcal{F}_T$ leads to a similar spectrum as δ so that only a scaling law links the two spectrum. The value $\chi_{\mathbf{0}}(0) = 1$ imply a scaling value of $\delta(0) = (2T + 1)/2\pi a$. Finally, $\mathbf{W}^{-1} = \delta(0)\mathbf{I}$ or

$$\mathbf{W}^{-1} = \frac{2T+1}{2\pi a}\mathbf{I},$$

where $\mathbf{I}$ denotes the identity operator of $\mathcal{F}_T$.

A similar conclusion occurs on the homogeneous sphere of triangular truncation T. In that case, the Dirac distribution δ corresponds to the distribution located at the North pole. Its spherical harmonic spectrum δ_n^m is $\delta_n^0 = Y_n^m(0)/4\pi a^2 = \sqrt{2n+1}/4\pi a^2$ (Y_n^m denotes the spherical harmonic function) when $m = 0$ and $\delta_n^m = 0$ elsewhere. Then it can be proofed that $\mathbf{W}^{-1} = \delta(0)\mathbf{I}$ with $\delta(0) = 1/4\pi a^2 \sum_{n=0}^{T}(2n + 1)$ or simply

$$\mathbf{W}^{-1} = \frac{(T+1)^2}{4\pi a^2}\mathbf{I},$$

where $\mathbf{I}$ denotes the identity operator of the subspace spanned by spherical harmonic within truncation T.

References

Belo Pereira M. and Berre L. 2006. *The use of an Ensemble approach to study the Background Error Covariances in a Global NWP model*. *Mon. Wea. Rev.*, **134**, 2466–2489.

Berre L., Pannekoucke O., Desrozier G., Stefanescu S., Chapnik B. and Raynaud L. 2007. *A variational assimilation ensemble and the spatial filtering of its error covariances: increase of sample size by local spatial averaging*. 151–168 in *Proceeding of workshop on flow-dependent aspects of data assimilation*, 11-13 June 2007. ECMWF, Reading, UK.

Boyd J.P. 2001. Chebyshev and Fourier Spectral Methods, Dover Publications (2 Revised edition).

Bouttier F. 1994. *A dynamical estimation of error covariances in an assimilation system*. *Mon. Wea. Rev.*, **122**, 2376–2390.

Buis S., Piacentini A., Déclat D. 2006. *PALM: A Computational framework for assembling high performance computing applications*. *Concurrency Computat.: Pract. Exper.*, **18**(2), 247–262.

Cariolle D. and Teyssèdre H. 2007. *A revised linear ozone photochemistry parameterization for use in transport and general circulation models: multi-annual simulations*. *Atmos. Chem. Phys.*, **7**, 2183–2196.

Courtier P., Andersson E., Heckley W., Pailleux J., Vasiljević D., Hamrud M., Hollingsworth A., Rabier F. and Fisher M. 1998. *The ECMWF implementation of three-dimensional variational assimilation (3D-Var). I: Formulation*. *Q.J.R. Meteorol. Soc.*, **124**, 1783–1807.

Daley R. 1991. *Atmospheric Data Analysis*. Cambridge University Press.

Dufour A., Amodei M., Ancellet G. and Peuch V.-H. 2004. *Observed and modelled "chemical weather" during ESCOMPTE*. *Atmos. Res.*, **74**, 161–189.

Egbert G.D., Bennett A.F. and Foreman M.G.G. 1994. *TOPEX/POSEIDON tides estimated using a global inverse model*. *J. Geophys. Res.* **99**, 24,821–24,852.

Fisher M. and Courtier P. 1995. *Estimating the covariance matrices of analysis and forecast error in variational data assimilation*. ECMWF Technical Memorandum, **220**.

Fisher M. and Anderson E. 2001. *Developments in 4D-Var and Kalman Filtering*. ECMWF Technical Memorandum, **347**, 38pp.

Fisher M. 2003. *Background error covariance modelling*. Processing of the ECMWF Seminar on "Recent developments in data assimilation for atmosphere and ocean", Reading, 8–12 September 2003, 45–63.

Gaspari G. and Cohn S. 1999. *Construction of correlation functions in two and three dimensions*. *Q.J.R. Meteorol. Soc.*, **125**, 723–757.

Geer A. J., Lahoz W. A., Bekki S., Bormann N., Errera Q., Eskes H. J., Fonteyn D., Jackson D. R., Juckes M. N., Massart S., Peuch V.-H., Rharmili S. and Segers A. 2006. *The ASSET intercomparison of ozone analyses: method and first results*. *Atmos. Chem. Phys.*, **6**, 5445–5474.

Gneiting T. 1999a. *Isotropic correlation functions on d-dimensional balls*. *Adv. Appl. Prob.*, **31**, 625–631.

Gneiting T. 1999b. *Correlation functions for atmospheric data analysis*. *Q.J.R. Meteorol. Soc.*, **125**, 2449–2464.

Houtekamer P.L., Lefaivre L., Derome J., Ritchie H. and Mitchell H.L. 1996. *A system simulation approach to ensemble prediction*. *Mon. Wea. Rev.*, **124**, 1225–1242.

Kalnay E. 2002. *Atmospheric modeling, data assimilation and predictability*. Cambridge University Press, p364.

Lahoz W. A., Geer A. J., Bekki S., Bormann N., Ceccherini S., Elbern H., Errera Q., Eskes H. J., Fonteyn D., Jackson D. R., Khattatov B., Marchand M., Massart S., Peuch V.-H., Rharmili S., Ridolfi M., Segers A., Talagrand O., Thornton H. E., Vik A. F. and von Clarmann T. 2007. *The Assimilation of Envisat data (ASSET) project*. *Atmospheric Chemistry and Physics*, **7**, 1773–1796.

Liu H., Xue M., Purser R.J., and Parrish D.F. 2007. *Retrieval of moisture from simulated GPS slant-path water vapor observations using 3DVAR with anisotropic recursive filters*. *Mon. Wea. Rev.*, **135**, 1506–1521.

Lorenc A. 1992. *Iterative analysis using covariance functions and filters*. *Q.J.R. Meteorol. Soc.*, **118**, 569–591.

Massart S., Cariolle D. and Peuch V.-H. 2005a. *Vers une meilleure représentation de la distribution et de la variabilité de l'ozone atmosphérique par l'assimilation des données satellitaires*. *C. R. Acad. Sci.*, **337**, 1305–1310.

Massart S., Manzoni H., Cariolle D., Peuch V.-H and Piacentini A. 2005b. *Validation of a 3D-Fgat assimilation of MIPAS ozone profiles in a global chemistry and transport model* in Proceeding of the Fourth WMO Symposium on Assimilation of Observations in Meteorology and Oceanography, Prague, Czech Republic, April 2005.

Massart S., Piacentini A., Cariolle D., El Amraoui L. and Semane N. 2007. *Assessment of the quality of the ozone measurements from the Odin/SMR instrument using data assimilation*. *Can. J. Phys.*, **85**, in press.

Pannekoucke O., Berre L. and Desroziers G. 2007. *Filtering properties of wavelets for local background-error correlations*. *Q.J.R. Meteorol. Soc.*, **133**, 363–379.

Pannekoucke O., Berre L. and Desroziers G. 2008. *Background error correlation length-scale estimates and their sampling statistics*. *Q.J.R. Meteorol. Soc.*, **134**, 497–508.

Pradier S., Attié J. L., Chong M., Escobar J., Peuch V. H., Lamarque J. F., Khattatov B. and Edwards D. 2006. *Evaluation of 2001 springtime CO transport over West Africa using MOPITT CO measurements assimilated in a global chemistry transport model*. *Tellus*, **58**, 163–176.

Purser R. J., Wu W.-S., Parrish D. and Roberts N. 2003a. *Numerical aspects of the application of recursive filters to variational statistical analysis. Part I: Spatially homogeneous and isotropic Gaussian covariances*. *Mon. Wea. Rev.*, **131**, 1524–1535.

Purser R. J., Wu W.-S., Parrish D. and Roberts N. 2003b. *Numerical aspects of the application of recursive filters to variational statistical analysis. Part II: Spatially inhomogeneous and anisotropic general covariances*. *Mon. Wea. Rev.*, **131**, 1536–1548.

Purser R.J., de Pondeca M., Parrish D. and Devenyi D. 2007. *Covariance modelling in a grid-point analysis*. 11–25 in *Proceeding of workshop on flow-dependent aspects of data assimilation*, 11-13 June 2007. ECMWF, Reading, UK.

Raspollini P., Belotti C., Burgess A., Carli B., Carlotti M., Ceccherini S., Dinelli B. M., Dudhia A., Flaud J.-M., Funke B., Höpfner M., López-Puertas M., Payne V., Piccolo C., Remedios J. J., Ridolfi M. and Spang R. 2006. *MIPAS level 2 operational analysis*. *Atmos. Chem. Phys.*, **6**, 5605–5630.

Teyssèdre H., Michou M., Clark H.L., Josse B., Karcher F., Olivié D., Peuch V.-H., Saint-Martin D., Cariolle D., Attié J.-L., Ricaud P., van der A R. J. and Chéroux F. 2007. *A new chemistry-climate tropospheric and stratospheric model MOCAGE-Climat: evaluation of the present-day climatology and sensitivity to surface processes*. *Atmos. Chem. Phys. Discuss.*, **7**, 11295-11398.

Weaver A. and Courtier P. 2001. *Correlation modelling on the sphere using a generalized diffusion equation*. *Quart. J. Roy. Meteor. Soc.*, **127**, 1815–1846.

Weaver A. and Ricci S. 2003 *Constructing a background-error correlation model using generalized diffusion operators*. In Proceedings of the ECMWF Seminar Series on Recent developments in atmospheric and ocean data assimilation, ECMWF, Reading, U. K., 8–12 September, pp. 327-340.

Weber R. and Talkner P. 1993. *Some remarks on spatial correlation models*. *Mon. Wea. Rev.*, **121**, 2611–2617.

